Information Integration
– Mock Exam –

Werner Nutt

04/06/12

- The exam comprises **five** questions, which consist of several subquestions.
- Each question is worth 15 points. The total mark for the test will be based on the **four** questions for which you achieved the highest mark.
- There is a total of 60 points that can be achieved in this exam. You will have 2 hours time to answer the questions.
- Please, write down the answers to your questions in the test booklet handed out to you.
- For drafts use the blank paper provided by the university.
- If the space in the booklet turns out to be insufficient, please use the university paper for additional answers and return them with the booklet.

Queries in Relational Algebra and Calculus

Suppose a boat club has a database with the schema

$$\texttt{Boat}(\texttt{bname}, \texttt{type}, \texttt{colour})$$
$$\texttt{Reservation}(\texttt{mname}, \texttt{bname}, \texttt{day})$$

which records information about the boats owned by the club and about which member has reserved which boat on which day.

Consider the following two queries:

1. "Which members have only reserved red boats? "

2. "Which members made reservations for every boat of type dinghy?"

Express each query

(i) in relational algebra

(ii) in relational calculus, that is, as an expression of the form

$$\{\, x \mid \phi(x) \,\}$$

where x is the variable for which we want bindings and $\phi(x)$ is a logical formula with free variable x.

Safety and Domain Independence of Queries

Consider the following four queries expressed in relational calculus:

1. $\{\, x, y \mid \exists z\, \texttt{hasChild}(x, z) \lor \exists w\, \texttt{hasChild}(w, y) \,\}$
2. $\{\, x \mid \texttt{rich}(x) \land \forall y\, (\texttt{hasChild}(x, y) \to \neg\texttt{rich}(y)) \,\}$
3. $\{\, x \mid \texttt{rich}(x) \land \forall y\, (\neg\texttt{hasChild}(x, y) \to \texttt{rich}(y)) \,\}$

For each query, determine whether or not it is

- safe
- domain-independent.

For each query and each property, if your answer is "yes", briefly and informally explain your answer. If your answer is "no", provide an example showing that the query does not have the property in question.

Containment

In this question, we only consider relational conjunctive queries, that is, queries that do not contain comparisons.

Suppose q_0 is a fixed conjunctive query.

- The **container problem** for q_0 is the following decision problem:

  Given a conjunctive query q, decide whether $q_0 \subseteq q$.

- The **containee problem** for q_0 is the following decision problem:

  Given a conjunctive query q, decide whether $q \subseteq q_0$.

Prove or disprove the following statements:

1. For every conjunctive query q_0, there is a polynomial time algorithm to decide the *container problem* for q_0.
2. For every conjunctive query q_0, there is a polynomial time algorithm to decide the *containee problem* for q_0.

To prove a statement, a sketch of an algorithm together with a short argument why it is polynomial is sufficient. To disprove the statement, find a query q_0 for which the problem in question is NP-hard. Again, a proof sketch is sufficient to show the NP-hardness.

1 NP-Hardness of Conjunctive Query Containment

We introduce the Balanced Complete Bipartite Subgraph problem, which is known to be NP-complete. First, some definitions:

- A (nondirected) graph $G = (V, E)$ is *bipartite* if V can be divided into two disjoint sets V_1, V_2 such that every edge connects a vertex in V_1 to a vertex in V_2. (In other words, there are no edges connecting vertices in V_1 or vertices in V_2.)
- The bipartite graph G is *balanced* if $|V_1| = |V_2|$. (In other words, the two parts of the graph have an equal number of nodes.)
- The bipartite graph G is *complete* if for all vertices $v_1 \in V_1$, $v_2 \in V_2$ there is an edge $e \in E$ such that e connects v_1 and v_2. (In other words, all vertices that possible can be connected, are connected.)

The Balanced Complete Bipartite Subgraph problem is the following:

Given: A bipartite graph G and a number $k > 0$.

Question: Does G have a balanced complete bipartite subgraph of size $2k$?

Recall that simple conjunctive queries have only relational atoms in their body, and no equalities or inequalities.

Show that containment of simple conjunctive queries is NP-hard by reducing the Balanced Complete Bipartite Subgraph Problem to Query Containment. Describe the reduction and briefly explain why it is correct.

Hint: Remember other reductions of problems that ask whether a graph contains a specific kind of pattern.

Translation of Queries

Suppose a library has a database with the schema

$$\begin{array}{l}\texttt{book}(\texttt{bookid}, \texttt{author}, \texttt{title}, \texttt{language}) \\ \texttt{borrows}(\texttt{reader}, \texttt{bookid}, \texttt{date}),\end{array}$$

which records information about books and about which reader has borrowed which books at which date.

(i) Consider the following query, expressed in relational algebra in the named perspective:

$$\pi_{\texttt{reader}}(\texttt{borrows}) \setminus \pi_{\texttt{reader}}(\texttt{borrows} \bowtie \sigma_{\texttt{language}='\texttt{English}'}(\texttt{book}))$$

Express the query equivalently in

- Relational Calculus (i.e., first order predicate logic)
- SQL without using the boolean operators `AND`, `OR`, `MINUS`, or `EXCEPT`.

(ii) Consider the following query, expressed using rules:

$$\begin{array}{lcl}\texttt{ans}(R) & \text{:–} & \texttt{borrows}(R, B1, D),\ \texttt{book}(B1, '\texttt{Dickens}', T1, L1), \\ & & \texttt{borrows}(R, B2, D),\ \texttt{book}(B2, '\texttt{Scott}', T2, L2) \\ \texttt{ans}(R) & \text{:–} & \texttt{borrows}(R, B1, D),\ \texttt{book}(B1, A1, T1, '\texttt{English}'), \\ & & \texttt{borrows}(R, B2, D),\ \texttt{book}(B2, A2, T2, '\texttt{French}')\end{array}$$

Express the query equivalently in

- Relational Algebra in the unnamed perspective
- SQL.